Representation Surgery

Theory and Practice of Affine Steering

Shashwat Singh*, Shauli Ravfogel*, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, Ponnurangam Kumaraguru

ICML 2024

Presented as a part of the Responsible and Safe AI course on NPTEL

Representations

LM's train to predict the next token.

LM's train to produce contextual representations (vectors for a sentence) to predict the next token.

Can we treat them as well-behaved multi-variate spaces for de-biasing and controlling generation?

Guardedness

An attribute **Z** is considered guarded if we **can't** classify along that attribute.

Eg: being unable to tell the gender of a noun based on the representations from it.

Affine Concept Erasure: an Affine transformation that guards a particular attribute.

$b(x) = Wx + b$

What do we want?

Make the vectors from a particular distribution look like those of another distribution.

(eg. make toxic generation vectors look like non-toxic vectors)

Leading to: guarding

We want to do this with the smallest change possible so as to preserve semantics unrelated to **Z**

he asked John to ... f*** off

leave immediately

Context and related work

Steering vectors :

$v \rightarrow h + v = h'$

$\mu_0 - \mu_1$

desirable

undesirable

Debiasing

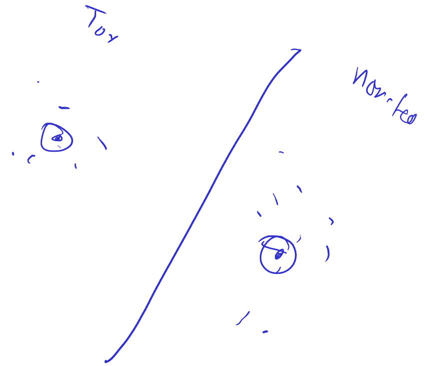
LEACE →

Guarding gender

asymmetry

Contribution 1

Existing literature uses steering vectors for this kind of thing. We provide a theoretical justification to steering vectors. We phrase an optimization problem for what we want



original

$$
\begin{aligned}
&\underset{s \in \mathrm{Aff}_s(D)}{\text{minimize}} \ \mathbb{E}\left[||\mathbf{H} - s(\mathbf{H})||_2^2 \right] \\
&\text{subject to } \ \mathbb{E}[s(\mathbf{H}_{\mathrm{c}})] = \mathbb{E}[s(\mathbf{H}_{\mathrm{c}'})]
\end{aligned}
$$

intervened

Affine: $Wh + b$

Piecewise: if ( ) h else ( ) $Wh + b$.

sentence, label available.

h_1, h_2, ..., h_4, ...

o $c \to \mu_c = E[h|c]$.

$c' \to \mu_{c'} = E[h|c']$

Contribution 1

Existing literature uses steering vectors for this kind of thing. We provide a theoretical justification to steering vectors. We phrase an optimization problem for what we want

$$\underset{s \in \mathrm{Aff}_s(D)}{\text{minimize}} \ \mathbb{E}\left[||\mathbf{H} - s(\mathbf{H})||_2^2\right]$$

$$\text{subject to } \ \mathbb{E}[s(\mathbf{H}_c)] = \mathbb{E}[s(\mathbf{H}_{c'})]$$

matching the first moment

Solution:

$$s^\star(\mathbf{H})(\boldsymbol{s}) = \begin{cases} \mathbf{H}(\boldsymbol{s}) + \boldsymbol{\mu}_{c'} - \mathbf{W}^\star \boldsymbol{\mu}_c & \textit{if} \quad \phi(\boldsymbol{s}) = c \\ \mathbf{H}(\boldsymbol{s}) & \textit{if} \quad \phi(\boldsymbol{s}) = c'. \end{cases}$$

$\mathbf{W}^\star = \mathbf{I}$

$\mu_{c'} - \mu_c \rightarrow$ undesirable.

$I \rightarrow$ desirable.

Contribution 2

x in some ways. → matching

→ 1 and 2 i.e match mean and covariance.

We extend the optimization problem and provide a more expressive steering function.

1st moment

$$\underset{s \in \mathrm{Aff}_s(D)}{\text{minimize}} \; \mathbb{E}\left[||\mathbf{H} - s(\mathbf{H})||_2^2 \right]$$

$$\text{subject to} \;\; \mathbb{E}[s(\mathbf{H}_c)] = \mathbb{E}[s(\mathbf{H}_{c'})]$$

$$\mathbb{E}[s(\mathbf{H}_c)s(\mathbf{H}_c)^\top] = \mathbb{E}[s(\mathbf{H}_{c'})s(\mathbf{H}_{c'})^\top]$$

2nd moment

Contribution 2

We extend the optimization problem and provide a more expressive steering function.

$$\underset{s \in \mathrm{Aff}_s(D)}{\text{minimize}} \; \mathbb{E}\left[||\mathbf{H} - s(\mathbf{H})||_2^2\right]$$

$$\text{subject to } \; \mathbb{E}[s(\mathbf{H}_c)] = \mathbb{E}[s(\mathbf{H}_{c'})]$$

$$\mathbb{E}[s(\mathbf{H}_c)s(\mathbf{H}_c)^\top] = \mathbb{E}[s(\mathbf{H}_{c'})s(\mathbf{H}_{c'})^\top]$$

Σ : covariance matrix.

has the solution

$$s^\star(\mathbf{H})(\boldsymbol{s}) = \begin{cases} \mathbf{W}^\star\mathbf{H}(\boldsymbol{s}) + \mathbf{b}^\star & \textit{if} \quad \phi(\boldsymbol{s}) = \mathrm{c} \\ \mathbf{H}(\boldsymbol{s}) & \textit{if} \quad \phi(\boldsymbol{s}) = \mathrm{c}'. \end{cases}$$

where we define

$$\mathbf{W}^\star = \boldsymbol{\Sigma}_\mathrm{c}{}^{-\frac{1}{2}}(\boldsymbol{\Sigma}_\mathrm{c}{}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\mathrm{c}'}\boldsymbol{\Sigma}_\mathrm{c}{}^{\frac{1}{2}})^{\frac{1}{2}}\boldsymbol{\Sigma}_\mathrm{c}{}^{-\frac{1}{2}}$$

$$\mathbf{b}^\star = -\mathbf{W}^\star\boldsymbol{\mu}_\mathrm{c} + \boldsymbol{\mu}_{\mathrm{c}'}.$$

(Optimal transport)

Implications

a no-gradient and cheap way to control generation and to de-bias.

De-biasing

Y = y

doctor

multiclas : classify professi-on

Guard : gender.

Protecting attributes without damage

Model	Intervention	TPR ↓	Accuracy ↑
BERT-base	Base	0.155	0.799
	LEACE	0.137	0.797
	Postprocessing (Xian et al., 2023)	0.146	0.742
	Mean Matching	0.141	0.797
	Mean+Covariance Matching	**0.093**	0.785
GPT-2	Base	0.168	0.676
	LEACE	0.093	0.670
	Postprocessing (Xian et al., 2023)	0.112	0.627
	Mean Matching	0.094	0.670
	Mean+Covariance Matching	**0.070**	0.660
Llama2-7b	Base	0.143	0.786
	LEACE	0.133	0.795
	Postprocessing (Xian et al., 2023)	-	-
	Mean Matching	0.139	0.797
	Mean+Covariance Matching	**0.085**	0.783

$$\text{TPR-Gap}(y) = \mathop{\mathbb{E}}_{\mathbf{h}_c \sim \mathbb{P}(\mathbf{H}_c | Y=y)} \mathbb{P}(\bar{Y} = y \mid \mathbf{H}_c = \mathbf{h}_c) - \mathop{\mathbb{E}}_{\mathbf{h}_{c'} \sim \mathbb{P}(\mathbf{H}_{c'} | Y=y)} \mathbb{P}(\bar{Y} = y \mid \mathbf{H}_{c'} = \mathbf{h}_{c'}).$$

$$\text{TPR}_{\text{RMS}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \text{TPR-Gap}(y_k)^2}.$$

Multi-class classification that should be unaffected by gender

Controlling generation

Perspective

Experiments:

Evaluating maximum toxicity of sentences

Model	Exp. Max. Tox. ↓	Tox. prob. ↓	Fluency ↓	1-gram ↑	2-gram ↑	3-gram ↑
GPT-2 (large)	0.39	0.25	24.66	0.58	0.85	0.85
DAPT	0.27	0.09	30.27	0.57	0.84	0.84
GeDI	0.24	0.06	48.12	0.62	0.84	0.83
PPLM (10%)	0.38	0.24	32.58	0.58	0.86	0.86
UDDIA	0.24	0.04	**26.83**	0.51	0.80	0.83
DExperts (large, all jigsaw)	0.21	**0.02**	27.15	0.56	0.84	0.84
GOODTRIEVER	0.22	0.04	27.11	0.58	0.82	0.83
Mean Matching	0.33	0.16	28.00	0.58	0.85	0.85
Mean+Covariance Matching	0.29	0.09	30.7	0.54	0.84	0.84

No fine-tuning control

Shashwat Singh

Shauli Ravfogel